Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Klasifikasi Pesan Penipuan pada Platform WhatsApp Menggunakan Metode Naïve Bayes Berbasis TF-IDF, N-Gram, dan Chi-Square

Hardika Nur Saputra, Ardytha Luthfiarta*

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email:[1]111202214747@mhs.dinus.ac.id, [2,*]ardytha.luthfiarta@dsn.dinus.ac.id
Email Penulis Korespondensi: *ardytha.luthfiarta@dsn.dinus.ac.id
Submitted: **13/05/2026**; Accepted: **22/06/2026**; Published**: 23/06/2026**

Abstrak−Pesatnya perkembangan komunikasi digital menyebabkan meningkatnya pertukaran pesan melalui berbagai platform yang diiringi dengan maraknya penyebaran pesan penipuan (*scam*). Kondisi tersebut menuntut adanya sistem otomatis yang mampu mengidentifikasi dan mengklasifikasikan pesan secara cepat dan akurat. Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi pesan pada platform WhatsApp berbasis teks menggunakan algoritma Naïve Bayes. Tahapan penelitian meliputi *preprocessing* teks yang terdiri dari *case folding, cleaning*, normalisasi, *stopword removal*, dan *stemming* untuk meningkatkan kualitas data. Selanjutnya, dilakukan ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang dikombinasikan dengan pendekatan N-Gram (*unigram*) untuk merepresentasikan setiap kata pada teks, serta diterapkan seleksi fitur Chi-Square untuk memperoleh fitur yang paling relevan dalam proses klasifikasi. Dataset yang digunakan terdiri dari tiga kategori pesan WhatsApp, yaitu normal, promosi, dan penipuan. Selain itu, penelitian ini juga menerapkan metode penyeimbangan data menggunakan *Random Oversampling* guna meningkatkan jumlah sampel kelas minoritas pada data latih agar performa model menjadi lebih optimal. Kontribusi utama penelitian ini adalah penerapan kombinasi TF-IDF unigram, seleksi fitur Chi-Square, dan Random Oversampling pada algoritma Naïve Bayes untuk meningkatkan performa klasifikasi pesan WhatsApp berbahasa Indonesia, khususnya pada kondisi distribusi kelas yang tidak seimbang. Evaluasi model dilakukan menggunakan *Confusion Matrix* dengan metrik *accuracy, precision, recall*, dan F1-score. Hasil pengujian menunjukkan bahwa model yang dibangun mampu mencapai tingkat akurasi sebesar 95,63%, sehingga metode yang digunakan terbukti efektif dalam mengklasifikasikan pesan WhatsApp secara akurat dan konsisten.

Kata Kunci: Klasifikasi Teks; Penipuan WhatsApp; Naive Bayes; TF-IDF; Chi-Square

Abstract−The rapid development of digital communication has led to an increase in message exchanges across various platforms, accompanied by the widespread spread of fraudulent messages (scams). This situation demands an automated system capable of identifying and classifying messages quickly and accurately. This study aims to develop a text-based message classification system on the WhatsApp platform using the Naïve Bayes algorithm. The research stages include text preprocessing consisting of case folding, cleaning, normalization, stopword removal, and stemming to improve data quality. Next, feature extraction is carried out using Term Frequency-Inverse Document Frequency (TF-IDF) combined with the N-Gram (unigram) approach to represent each word in the text, and Chi-Square feature selection is applied to obtain the most relevant features in the classification process. The dataset used consists of three categories of WhatsApp messages: normal, promotional, and fraudulent. In addition, this study also applies a data balancing method using Random Oversampling to increase the number of minority class samples in the training data for optimal model performance. The main contribution of this research is the application of a combination of TF-IDF unigram, Chi-Square feature selection, and Random Oversampling in the Naïve Bayes algorithm to improve the classification performance of Indonesian WhatsApp messages, especially in conditions of unbalanced class distribution. Model evaluation is carried out using a Confusion Matrix with accuracy, precision, recall, and F1-score metrics. The test results show that the model built is able to achieve an accuracy level of 95.63%, so the method used is proven to be effective in classifying WhatsApp messages accurately and consistently.

Keywords: Text Classification; Scam WhatsApp; Naive Bayes; TF-IDF; Chi-Square

1. PENDAHULUAN

Pesan digital pada platform WhatsApp saat ini menjadi salah satu alat komunikasi yang paling sering digunakan untuk saling bertukar pesan. Perkembangan teknologi komunikasi digital memberikan kemudahan bagi masyarakat untuk berinteraksi secara cepat dan praktis[1]. Kemudahan akses dan keefisien waktu menjadi alasan meningkatnya pengguna untuk tiap harinya[2]. Hal tersebut disebabkan karena munculnya teknologi-teknologi dibidang komunikasi[3]. Di sisi lain, kenaikan tersebut disalahgunakan oleh oknum kriminal dengan menjadikan peluang bagi mereka untuk melakukan tindak kejahatan yaitu penipuan melalui media komunikasi digital[4]. Akses informasi saat ini juga semakin mudah serta lebih cepat diakses. Pesan Penipuan dalam bentuk pesan digital ini menjadi isu yang semakin memprihatinkan. Pesan penipuan seringkali mengecoh penerima dengan menyamar sebagai komunikasi resmi, sehingga perlu keahlian untuk mengidentifikasinya [4]. Oleh karena itu, klasifikasi pesan penipuan ini menjadi topik yang penting dalam penelitian keamanan siber dan pengolahan bahasa alami (Natural Language Processing/NLP)[5].

Kementerian Komunikasi dan Informatika telah berupaya melawan kejahatan cyber, termasuk penipuan online, dengan memblokir ribuan situs berita palsu yang ada di indonesia [4]. Tantangan semakin kompleks dengan munculnya pesan penipuan digital yang diterima para pengguna, seperti pesan penipuan transfer uang, pulsa, hadiah dan undian palsu atau tindakan kriminal lainnya [4]. Kemkominfo mengidentifikasi sekitar 800.000 situs berita palsu yang tersebar di berbagai wilayah, dan pada tahun 2016 tercatat telah memblokir sebanyak 773.000 situs web dalam 10 kategori termasuk pornografi, narkoba, judi, isu SARA, penipuan, radikalisme, kekerasan anak, keamanan internet,

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

hingga pelanggaran hak kekayaan intelektual[4]. Data statistik periode Agustus 2018 sampai Februari 2023 menunjukkan adanya 1.730 kasus penipuan daring yang teridentifikasi. Total kerugian finansial yang diakibatkan oleh aktivitas ilegal tersebut diestimasikan mencapai Rp18,7 triliun [6]. Berdasarkan data tersebut, banyaknya angka yang terjadi ini mencerminkan dampak serius yang ditimbulkan oleh kejahatan cyber dalam bentuk penipuan online terhadap masyarakat dan juga ekonomi Indonesia [6]. Dengan demikian, diperlukan sistem deteksi pesan penipuan secara otomatis agar pengguna dapat lebih mudah mengenali pesan yang mencurigakan serta mengurangi risiko terjadinya penipuan digital yang semakin banyak ditemukan pada berbagai platform komunikasi.[4].

Sistem deteksi pesan penipuan ini dibangun dengan mengombinasikan metode Naïve Bayes serta Teknik fitur *TF-IDF* dan *N-Gram unigram.* TF-IDF digunakan untuk menganalisis teks dan menilai kontribusi setiap kata berdasarkan relevansinya terhadap dokumen, sedangkan N-Gram digunakan untuk menangkap pola urutan kata yang dapat meningkatkan representasi konteks dalam teks, Selanjutnya, metode *Chi-Square* diterapkan sebagai Teknik seleksi fitur untuk memilih fitur-fitur yang paling relevan dan berpengaruh terhadap proses klasifikasi. Algoritma *Naïve Bayes* kemudian memanfaatkan fitur-fitur tersebut untuk mengidentifikasi pola dan mengklasifikasikan pesan ke dalam kategori seperti penipuan, promosi, atau pesan normal. Selain itu, Sistem ini dikembangkan dengan kemampuan deteksi *real-time* untuk memastikan setiap pesan dikategorikan secara tepat dan akurat. Implementasi ini mengutamakan efektivitas klasifikasi serta aksesibilitas bagi pengguna tanpa memerlukan prosedur yang kompleks.

Beberapa penelitian sebelumnya telah mengkaji penerapan Algoritma Naive Bayes dalam klasifikasi teks dengan memanfaatkan teknik ekstraksi fitur seperti TF-IDF. Pendekatan tersebut menunjukkan kinerja yang cukup baik dalam berbagai kasus, termasuk deteksi spam dan pengolahan data teks tidak terstruktur[4]. Namun, sebagian penelitian masih berfokus pada penggunaan satu jenis metode ekstraksi fitur, sehingga belum mengoptimalkan potensi peningkatan akurasi melalui kombinasi beberapa teknik [4].

Penelitian mengenai klasifikasi pesan spam dan penipuan telah banyak dilakukan menggunakan berbagai metode machine learning. Penelitian oleh Sanhaji pada tahun 2024 menggunakan metode TF-IDF dan Naïve Bayes dengan akurasi sebesar 91% [4]. Selanjutnya, penelitian oleh Pradana pada tahun 2025 menerapkan metode TF-IDF, Chi-Square, dan Naïve Bayes dengan akurasi sebesar 93%[7]. Penelitian oleh Dwiprayoga pada tahun 2025 menggunakan metode BoW, TF-IDF, dan Naïve Bayes dengan akurasi sebesar 94,44% [8]. Selain itu, penelitian oleh Pratama dan Ardianto pada tahun 2025 menggunakan metode TF-IDF dan Support Vector Machine (SVM) untuk klasifikasi SMS spam berbahasa Indonesia yang terdiri dari kategori normal, penipuan, dan promosi dengan akurasi sebesar 92%[9]. Berdasarkan beberapa penelitian tersebut, penelitian ini mengusulkan kombinasi Random Oversampling, TF-IDF unigram, Chi-Square, dan Naïve Bayes yang mampu memperoleh akurasi sebesar 95,63%.

Berdasarkan beberapa penelitian terdahulu, masih terdapat beberapa celah penelitian yang perlu dikaji lebih lanjut. Penelitian oleh Sanhaji menunjukkan bahwa kombinasi TF-IDF dan Naïve Bayes mampu menghasilkan performa klasifikasi yang baik, namun belum memanfaatkan teknik seleksi fitur untuk mengurangi fitur yang kurang relevan. Penelitian Pradana telah menerapkan Chi-Square sebagai seleksi fitur, tetapi belum mengombinasikannya dengan pendekatan N-Gram untuk menangkap pola kata yang lebih representatif. Penelitian Dwiprayoga menggunakan BoW dan TF-IDF dalam proses klasifikasi, namun belum membahas penanganan ketidakseimbangan distribusi kelas yang dapat memengaruhi performa model. Selain itu, penelitian oleh Pratama dan Ardianto menggunakan metode TF-IDF dan Support Vector Machine (SVM) untuk mendeteksi SMS spam berbahasa Indonesia dengan akurasi sebesar 92%, namun penelitian tersebut berfokus pada platform SMS dan belum mengkaji klasifikasi pesan penipuan pada platform WhatsApp yang memiliki karakteristik bahasa, pola komunikasi, serta variasi pesan yang berbeda. Penelitian tersebut juga belum menerapkan teknik penyeimbangan data maupun analisis kontribusi setiap tahapan pemrosesan data terhadap hasil klasifikasi. Oleh karena itu, penelitian ini mengusulkan kombinasi Random Oversampling, TF-IDF unigram, seleksi fitur Chi-Square, dan algoritma Naïve Bayes untuk mengatasi ketidakseimbangan data, meningkatkan kualitas representasi fitur, serta menghasilkan performa klasifikasi pesan penipuan pada platform WhatsApp berbahasa Indonesia yang lebih optimal.

Dengan demikian, terdapat tiga celah penelitian utama yang belum banyak dikaji pada penelitian sebelumnya, yaitu belum adanya kombinasi TF-IDF unigram, seleksi fitur Chi-Square, dan penanganan ketidakseimbangan data menggunakan Random Oversampling pada klasifikasi pesan penipuan platform WhatsApp berbahasa Indonesia. Selain itu, penelitian terdahulu juga belum melakukan evaluasi komprehensif terhadap kontribusi masing-masing tahapan pemrosesan data terhadap peningkatan performa klasifikasi.

Merujuk pada celah penelitian yang telah diidentifikasi, penelitian ini memberikan beberapa kontribusi utama. Pertama, penelitian ini mengintegrasikan teknik ekstraksi fitur TF-IDF unigram, seleksi fitur Chi-Square, dan algoritma Naïve Bayes dalam klasifikasi pesan penipuan berbahasa Indonesia. Kedua, penelitian ini menerapkan metode Random Oversampling untuk mengatasi permasalahan ketidakseimbangan distribusi kelas pada dataset sehingga proses pelatihan model menjadi lebih seimbang. Ketiga, penelitian ini melakukan evaluasi komparatif terhadap beberapa skenario pemodelan, mulai dari penggunaan Naïve Bayes dasar hingga kombinasi TF-IDF, Chi-Square, dan Random Oversampling untuk mengetahui konfigurasi terbaik. Keempat, penelitian ini menghasilkan model klasifikasi yang mampu mencapai akurasi sebesar 95,63% sehingga menunjukkan bahwa pendekatan yang diusulkan dapat digunakan sebagai alternatif yang efektif dan efisien dalam mendeteksi pesan penipuan digital berbahasa Indonesia.

Selain itu, perkembangan teknologi yang semakin pesat menyebabkan penyebaran pesan digital menjadi lebih luas dan sulit dikendalikan. Hal ini membuka peluang bagi pelaku kejahatan untuk menyebarkan berbagai bentuk

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

penipuan dengan teknik yang semakin beragam dan sulit dikenali secara manual. Oleh karena itu, diperlukan suatu sistem otomatis yang mampu mendeteksi pola-pola pesan penipuan secara cepat dan akurat. Penggunaan pendekatan berbasis *Machine Learning* menjadi solusi yang relevan karena mampu mempelajari karakteristik data teks dan melakukan klasifikasi secara efisien.

Di tengah perkembangan metode Deep Learning seperti BERT dan Transformer yang banyak digunakan pada klasifikasi teks, algoritma Naïve Bayes masih memiliki keunggulan dari sisi efisiensi komputasi dan kecepatan proses pelatihan. Metode tersebut juga tetap efektif digunakan pada dataset teks berukuran terbatas. Oleh karena itu, penelitian ini menggunakan kombinasi TF-IDF, N-Gram unigram, dan seleksi fitur Chi-Square untuk meningkatkan representasi fitur teks sehingga model mampu mengenali pola pesan penipuan secara lebih optimal tanpa memerlukan sumber daya komputasi yang kompleks.

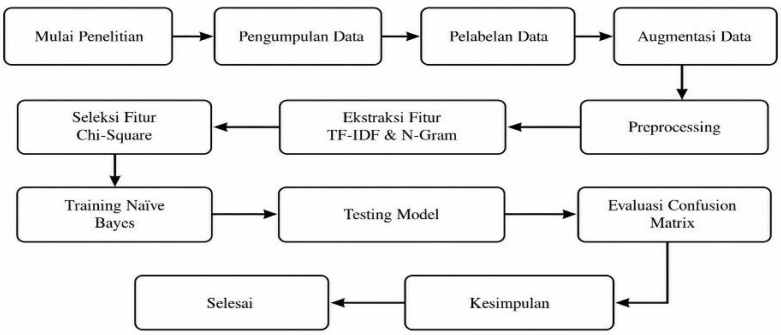
Meskipun berbagai penelitian telah menunjukkan bahwa Algoritma Naïve Bayes memiliki performa yang cukup baik dalam proses klasifikasi teks, namun masih terdapat beberapa kekurangan yang perlu dikembangkan lebih lanjut, seperti belum optimalnya kombinasi teknik ekstraksi dan seleksi fitur dalam meningkatkan akurasi model [4]. Selain itu, sebagian besar penelitian masih berfokus pada proses klasifikasi tanpa mengintegrasikan sistem ke dalam media yang dapat digunakan secara langsung oleh pengguna [10]. Melalui penggabungan teknik TF-IDF, N-Gram unigram, dan seleksi fitur Chi-Square, penelitian ini bertujuan untuk membangun sistem pemodelan yang lebih presisi. Pendekatan ini diharapkan mampu meningkatkan akurasi klasifikasi sekaligus menawarkan performa komputasi yang responsif dalam mengenali pola pesan penipuan pada platform WhatsApp.

Studi ini mengusulkan suatu kerangka klasifikasi pesan penipuan pada platform WhatsApp berbahasa Indonesia menggunakan kombinasi TF-IDF unigram, seleksi fitur Chi-Square, Random Oversampling, dan algoritma Naïve Bayes. Pendekatan yang diusulkan bertujuan untuk meningkatkan kualitas representasi fitur, mengatasi ketidakseimbangan data, serta menghasilkan model klasifikasi yang lebih akurat dan efisien dalam mengidentifikasi pesan penipuan secara otomatis.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan sistematis yang bertujuan untuk mengubah data pesan digital tak terstruktur menjadi pengklasifikasian pesan penipuan yang presisi. Adapun prosedur kerjanya mencakup pengumpulan data, praproses teks, pembobotan fitur melalui kombinasi TF-IDF dan N-Gram unigram, penyaringan fitur dengan metode Chi-Square, hingga tahap klasifikasi yang mengimplementasikan algoritma Naive Bayes. Alur kerja penelitian ini secara visual dapat dilihat pada Gambar 1.



Gambar 1. Alur Tahapan Penelitian

Gambar 1 menunjukkan alur pengolahan data dalam penelitian ini yang dimulai dari tahap pengumpulan dataset publik pesan digital (dataset_sms_spam_v1.csv & key_norm.csv), yang terdiri dari 1143 data dengan dua kolom utama yaitu kolom teks pesan yang berisi isi pesan dalam bentuk teks dan kolom label yang mengklasifikasikan pesan ke dalam tiga kategori, yaitu label 0 untuk pesan normal, label 1 untuk pesan penipuan, dan label 2 untuk pesan promosi. Selain itu, file key_norm.csv digunakan sebagai kamus normalisasi kata yang berisi dua kolom, yaitu kolom singkatan dan kolom hasil normalisasi, misalnya kata “yg” diubah menjadi “yang”, sehingga membantu proses normalisasi teks pada tahap preprocessing. Setelah proses pengumpulan dan pelabelan data, dilakukan tahap penyeimbangan data menggunakan metode Random Oversampling untuk mengatasi ketidakseimbangan distribusi kelas pada dataset. Metode ini dilakukan dengan menambahkan sampel pada kelas minoritas hingga jumlah data pada setiap kategori menjadi lebih seimbang, sehingga model dapat mempelajari pola dari setiap kelas secara lebih optimal dan mengurangi potensi bias terhadap kelas mayoritas. Selanjutnya dilakukan tahap preprocessing teks yang meliputi proses case folding, normalisasi teks, stopword removal, dan stemming untuk membersihkan serta menormalkan data agar siap diproses lebih lanjut. Tahap selanjutnya adalah proses ekstraksi fitur menggunakan metode TF-IDF dan N-

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Gram unigram guna mengonversi data teks ke dalam format numerik guna memenuhi persyaratan input bagi algoritma klasifikasi.

Selanjutnya, diterapkan metode *Chi-Square* sebagai teknik seleksi fitur untuk menentukan fitur yang memiliki tingkat keterkaitan dan kontribusi paling tinggi dalam proses klasifikasi.[7]. Tahap klasifikasi selanjutnya dilakukan menggunakan algoritma Naïve Bayes untuk mengidentifikasi dan mengategorikan pesan ke dalam kelas penipuan, promosi, maupun pesan normal[1]. Tahap berikutnya adalah evaluasi model menggunakan Confusion Matrix untuk mengevaluasi kinerja sistem melalui perhitungan nilai *accuracy, precision, dan recall*.

Selain itu, model yang dikembangkan dirancang dengan arsitektur yang efisien sehingga mampu memproses dan mendeteksi pesan secara instan. Fokus pada optimalisasi model ini diharapkan dapat membantu identifikasi dini terhadap potensi penipuan, guna meminimalisasi risiko kerugian akibat kejahatan digital yang kian meningkat.

2.2 Pengumpulan Data

Data yang menjadi basis analisis dalam penelitian ini adalah kumpulan pesan digital yang diperoleh dari dataset publik terkait klasifikasi spam dan pesan penipuan [8]. Dataset utama yang digunakan berasal dari repositori https://github.com/ksnugroho/klasifikasi-spam-sms yang berisi data pesan digital dalam beberapa kategori, yaitu pesan normal, pesan promosi, dan pesan penipuan [11]. Dataset terdiri dari teks pesan serta label klasifikasi yang digunakan sebagai data latih dan data uji pada proses pemodelan. Penggunaan dataset publik bertujuan untuk memperoleh variasi pola pesan yang beragam sehingga model yang dibangun mampu mengenali karakteristik pesan secara lebih akurat. Selain itu, penelitian ini juga menggunakan file key_norm.csv sebagai kamus normalisasi kata tidak baku untuk mendukung proses preprocessing teks [11].

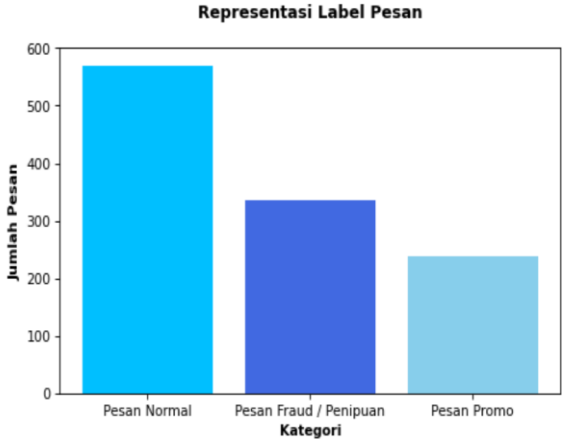
Tabel 1. Dataset Penelitian

No	Jenis Pesan	Jumlah
1	Pesan Normal	569
2	Pesan Penipuan	335
3	Pesan Promosi	239

Tabel 2. Dataset Kamus Normalisasi

Singkatan	Hasil
yg	yang
bbrpa	beberapa
bpk	bapak
dah	sudah
dateng	datang
ama	sama
bntr	bentar
bole	boleh

Berdasarkan Tabel 1, dataset penelitian terdiri dari 1.143 data pesan digital yang terbagi ke dalam tiga kategori, yaitu 569 pesan normal, 335 pesan penipuan, dan 239 pesan promosi. Pembagian kategori tersebut digunakan untuk melatih model klasifikasi agar mampu mengenali karakteristik masing-masing jenis pesan secara lebih akurat. Sementara itu, Tabel 2 menunjukkan kamus normalisasi yang digunakan pada tahap preprocessing untuk mengubah kata tidak baku, singkatan, atau bahasa informal menjadi kata baku sesuai kaidah bahasa Indonesia. Tujuan normalisasi adalah menyeragamkan kata agar variasi yang tidak perlu berkurang, sehingga representasi teks menjadi lebih baik dan proses klasifikasi pesan digital lebih efektif..



Gambar 2. Jumlah Dataset Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Berdasarkan Gambar 4, distribusi data menunjukkan bahwa kategori pesan normal memiliki jumlah data terbanyak, yaitu 569 data, diikuti pesan penipuan sebanyak 335 data, dan pesan promo sebanyak 239 data. Perbedaan jumlah data pada setiap kategori menunjukkan adanya ketidakseimbangan kelas pada dataset penelitian

2.3 Pelabelan Data

Pelabelan data bertujuan untuk mengubah data pesan digital yang masih mentah menjadi kategori tersetruktur agar dapat diproses oleh system klasifikasi. Dalam penelitian ini, proses pelabelan dilakukan dengan mengelompokkan setiap pesan ke dalam tiga kategori utama, yaitu pesan penipuan, pesan promosi, dan pesan normal. Kategori penipuan mencakup pesan yang mengandung indikasi kejahatan seperti permintaan transfer uang, hadiah paslu, atau penyamaran identitas. Kategori promosi berisi pesan yang berkaitan dengan penawaran produk atau jasa, seperti iklan dan diskon. Sementara itu, kategori normal merupakan pesan komunikasisehari-hari yang tidak mengandung unsur penipuan maupun promosi.

Tabel 3. Aturan Pelabelan Kategori Pesan

Kategori Pesan	Label	Deskripsi
Penipuan	0	Pesan yang mengandung unsur penipuan seperti permintaan transfer uang, hadiah palsu, atau penyemaran identitas
Promosi	1	Pesan yang berisi penawaran produk/jasa, diskon, atau iklan tanpa unsur penipuan
Normal	2	Pesan komunikasi umum yang tidak termasuk dalam kategori penipuan maupun promosi

2.4 Random Oversampling

Tahap Random Oversampling dilakukan setelah proses pelabelan data untuk menyeimbangkan distribusi jumlah data pada setiap kategori pesan. Pada penelitian ini, teknik Random Oversampling diterapkan dengan menambahkan data pada kelas minoritas secara acak hingga jumlah data setiap kategori menjadi lebih seimbang. Proses ini bertujuan untuk mengurangi ketidakseimbangan kelas yang dapat memengaruhi performa model klasifikasi. Dengan distribusi data yang lebih merata, model diharapkan mampu mempelajari pola pada setiap kategori pesan secara lebih optimal sehingga meningkatkan kemampuan generalisasi dan performa klasifikasi secara keseluruhan.

2.4 Pre-processing Data

Tahap *preprocessing* merupakan proses penting dalam pengolahan data teks untuk membersihkan data pesan digital dari berbagai gangguan (noise) agar siap diproses oleh algoritma Naïve Bayes. Mengingat pesan digital umumnya ditulis dengan bahasa informal, singkatan, serta penggunaan simbol yang beragam, maka proses prapemrosesan yang konsisten sangat diperlukan untuk meningkatkan kualitas fitur teks. Tahapan ini bertujuan untuk menyeragamkan struktur data sehingga sistem dapat mengenali pola kata dengan lebih akurat dalam proses klasifikasi pesan penipuan. Langkah-langkah prapemrosesan yang dilakukan dalam penelitian ini tercantum dalam poin-poin utama yaitu:

a. Case Folding: Pada tahapan ini, seluruh teks dalam pesan diseragamkan dengan mentransformasikan karakter menjadi huruf kecil secara keseluruhan[3], (*lowercase*) untuk menyamakan bentuk penulisan teks serta mengurangi perbedaan representasi kata yang disebabkan oleh penggunaan huruf besar dan huruf kecil.
b. Normalization (Normalisasi): Mengubah kata-kata tidak baku, istilah gaul *(slang),* atau singkatan yang sering muncul dalam dataset (seperti "bgttt", "gacocok", "recom") menjadi kata baku sesuai Ejaan Yang Disempurnakan (EYD). Kualitas hasil dari tahapan ini sangat menentukan performa klasifikasi karena teks yang bersih akan mengurangi dimensi fitur yang tidak relevan dalam perhitungan probabilitas
c. Filtering (Stopword Removal): Melakukan eliminasi terhadap kata-kata yang sering digunakan tetapi tidak memberikan informasi penting maupun pengaruh sentimen yang signifikan, seperti kata sandang, kata hubung, dan kata ganti (contoh: "yang", "dan", "saya")[12]. menghilangkan kata-kata umum yang tidak terlalu berpengaruh pada teks [13].
d. Stemming: Tahap akhir di mana setiap kata dikembalikan ke bentuk dasarnya dengan menghilangkan imbuhan (awalan, sisipan, atau akhiran). Proses tersebut memerlukan kumpulan kata dasar sebagai acuan, yang diperoleh dari Kamus Besar Bahasa Indonesia (KBBI) [14]. Dalam penelitian ini, digunakan algoritma Sastrawi untuk memproses teks bahasa Indonesia, sehingga kata-kata seperti "melembabkan" dan "kelembaban" akan disatukan menjadi kata dasar "lembab"[15]. Dengan standarisasi kata melalui seluruh rangkaian ini, algoritma dapat lebih mudah memetakan hubungan antara kata kunci tertentu dengan label sentimen yang telah ditentukan secara konsisten [12].

2.5 Exstraksi Fitur TF-IDF dan N-Gram

Tahap exstraksi fitur bertujuan untuk mengubah data teks pesan yang telah melalui proses preprocessing menjadi bentuk numerik agar dapat diolah oleh algoritma Naïve Bayes [16]. Fitur asli akan melalui proses pemetaan fungsional pada tahap feature extraction yang akan menghasilkan fitur baru [17]. Implementasi TF-IDF diawali dengan proses tokenisasi untuk menentukan frekuensi kemunculan setiap kata unik dalam suatu dokumen. Tahapan selanjutnya melibatkan penghitungan nilai *Term Frequency* (TF) yang dinormalisasi berdasarkan total kata, diikuti dengan penentuan bobot *Inverse Document Frequency* (IDF). Hasil akhirnya disajikan dalam format tabel, di mana bobot TF-

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

IDF diperoleh melalui hasil perkalian antara variabel TF dan IDF [10]. Pada penelitian ini digunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency)* untuk menghitung bobot setiap kata berdasarkan frekuensi kemunculannya dalam dokumen serta distribusinya pada keseluruhan data. Nilai *Term Frequency* merepresentasikan intensitas kemunculan suatu kata dalam sebuah pesan, sementara *Inverse Document Frequency* berperan dalam memberikan bobot lebih pada kata-kata langka yang dipandang memiliki nilai informatif lebih tinggi. Selain itu pendekatan N-Gram juga digunakan untuk menangkap pola urutan kata dalam teks, sehingga system dapat mengenali kombinasi kata yang sering muncul pada pesan penipaun. Penggunaan N-Gram membantu meningkatkan representasi konteks dibandingkan hanya menggunakan kata Tungga l[18]. Pendekatan N-Gram yang diterapkan pada penelitian ini secara spesifik menggunakan konfigurasi *unigram*, di mana teks dipecah menjadi kata tunggal mandiri guna menjaga kesederhanaan fitur namun tetap efektif dalam merepresentasikan bobot informasi setiap kosakata.Hasil dari proses ekstraksi fitur ini berupa matriks numerik yang selanjutnya digunakan sebagai input dalam proses pelatihan dan klasifikasi menggunakan algoritma Naïve Bayes [19].

2.6 Naïve Bayes

Algoritma Naïve Bayes adalah metode klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen satu sama lain [12]. Dalam penelitian ini, Algoritma digunakan untuk mengklasifikasikan pesan digital ke dalam beberapa kategori, yaitu penipuan, promosi, dan pesan normal, berdasarkan pola kata yang terdapat dalam teks pesan. Metode ini bekerja dengan menghitung probabilitas kemunculan kata-kata tertentu pada masing-masing kategori, sehingga sistem dapat menentukan kemungkinan suatu pesan termasuk ke dalam kelas tertentu. Naïve Bayes dipilih karena memiliki kemampuan yang baik dalam kelas tertentu. Naïve Bayes dipilih karena memiliki kemampuan yang baik dalam menangani data teks berukuran besar, proses komputasi yang relative cepat, serta performa yang cukup tinggi dalam tugas klasifikasi teks [12]. Secara matematis, Teorema Bayes dinyatakan sebagai berikut [20].

$$P(C_k \mid d) = \frac{P(C_k) \prod_{i=1}^{n} P(w_i|C_k)}{P(d)} \quad (1)$$

Berdasarkan Rumus 1, $P(C_k \mid d)$merupakan probabilitas posterior yang menunjukkan peluang sebuah dokumen atau pesan $d$termasuk ke dalam kelas C_k. Nilai $P(C_k)$adalah probabilitas prior dari masing-masing kelas sebelum dilakukan proses klasifikasi. Selanjutnya, $P(w_i \mid C_k)$menunjukkan probabilitas kemunculan kata ke-i pada kelas tertentu, sedangkan simbol $\prod$menyatakan perkalian seluruh probabilitas kata dalam dokumen. Adapun $P(d)$merupakan probabilitas dari dokumen berperan dalam nilai normalisasi. Pada studi ini, algoritma Naïve Bayes memanfaatkan representasi fitur hasil ekstraksi TF-IDF dan N-Gram unigram untuk mempelajari pola kata pada setiap kategori pesan sehingga sistem dapat melakukan klasifikasi pesan digital secara otomatis.

2.8 Chi-Square

Chi-Square merupakan metode seleksi fitur yang digunakan untuk mengukur tingkat keterkaitan antara fitur (kata) dengan kelas tertentu dalam proses klasifikasi. Pada penelitian ini, Metode Chi-Square diterapkan untuk menyeleksi fitur yang memiliki relevansi dan pengaruh paling signifikan terhadap kategori pesan. Melalui reduksi dimensi data ini, efisiensi komputasi serta tingkat akurasi model dapat ditingkatkan secara optimal. Proses seleksi dilakukan dengan menghitung nilai Chi-Square dari setiap fitur terhadap masing-masing kelas, kemudian memilih fitur dengan nilai tertinggi sebagai fitur yang paling informatif. Dengan demikian, fitur yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi akan dieliminasi, sehingga model Naïve Bayes dapat bekerja lebih optimal dalam mengklasifikasikan pesan digital.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

Berdasarkan Rumus (2), metode Chi-Square digunakan untuk mengukur tingkat hubungan antara suatu fitur dengan kategori kelas tertentu pada proses klasifikasi. Simbol χ^2menyatakan nilai Chi-Square, sedangkan O (*observed value*) merupakan nilai observasi aktual atau jumlah kemunculan fitur pada data asli. Sementara itu, E adalah nilai harapan (*expected value*) yang diperoleh berdasarkan distribusi data. Proses perhitungan dilakukan dengan menghitung selisih antara nilai observasi dan nilai harapan, kemudian dikuadratkan dan dibagi dengan nilai harapan. Semakin besar nilai Chi-Square yang dihasilkan, maka semakin tinggi tingkat relevansi fitur terhadap kategori kelas tertentu. Oleh karena itu, metode Chi-Square bertujuan untuk mengidentifikasi fitur dengan kontribusi paling signifikan terhadap model. Dengan menyaring fitur-fitur tersebut, proses klasifikasi pesan digital dapat dilakukan secara lebih optimal, baik dari aspek performa maupun penggunaan sumber daya.

2.9 Pengujian Model

Pengujian model merupakan tahap akhir untuk mengetahui tingkat kemampuan algoritma Naïve Bayes dalam mengelompokkan pesan digital secara tepat.[12]. Dalam penelitian ini, pengujian dilakukan menggunakan *Confusion Matrix*, yaitu alat evaluasi yang membandingkan hasil prediksi sistem dengan label aktual pada data uji[12]. Metrik utama yang digunakan meliputi *accuracy* untuk melihat tingkat ketepatan hasil prediksi, *precision* untuk mengukur ketepatan sistem dalam mengklasifikasikan kategori pesan tertentu, serta *recall* untuk mengukur kemampuan sistem

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

dalam menemukan kembali data yang sesuai dengan kategori sebenarnya. Proses pengujian dilakukan dengan membagi dataset menjadi data latih (*training*) dan data uji (*testing*). Evaluasi performa melalui *Confusion Matrix* sangat penting karena memungkinkan peneliti mengidentifikasi kesalahan klasifikasi, baik berupa *False Positive* maupun *False Negative*. Dengan menggunakan parameter evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*, kualitas model yang dibangun dapat divalidasi secara objektif untuk memastikan bahwa sistem deteksi pesan penipuan memiliki performa yang baik dalam mengelompokkan pesan ke dalam kategori penipuan, promosi, dan pesan normal.

3.0 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa algoritma Naïve Bayes dalam mengklasifikasikan pesan digital ke dalam kategori yang tepat, yaitu penipuan, promosi, dan pesan normal. Instrumen utama yang digunakan adalah *Confusion Matrix*, yang menyajikan perbandingan antara hasil prediksi sistem dengan data aktual pada dataset pengujian. Deskripsi dari setiap kategori dalam matriks evaluasi tersebut dapat dilihat pada Tabel 5.

Tabel 4. Deskripsi Komponen Confuison Matrix

Komponen	Nama	Deskripsi Operasional (Prsan Digital)
TP	True Positive	Kondisi di mana pesan aktual merupakan pesan penipuan dan sistem berhasil menprediksinya sebagai penipuan dengan benar.
TN	True Negative	Kondisi di mana pesan aktual bukan penipuan (promosi atau normal) dan sistem berhasil mengklasifikasikannya dengan benar.
FP	False Positive	Kondisi di mana pesan aktual bukan penipuan namun sistem salah memprediksinya sebagai pesan penipuan.
FN	False Negative	Kondisi di mana pesan aktual merupakan pesan penipuan namun sistem gagal mendeteksinya dan mengklasifikasikannya sebagai kategori lain.

Setelah mendapatkan nilai dari *Confusion Matrix*, dilakukan perhitungan metrik evaluasi guna menilai performa model klasifikasi secara kuantitatif.. Penjelasan mengenai metrik evaluasi tersebut adalah sebagai berikut:

a. Accuracy (Akurasi):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Accuracy yang didapat berdasarkan empat poin pemetaan didapat menggunakan persamaan rumus (3). Perlu dicatat, bahwa accuracy bukan suatu tolak ukur yang bagus jika data yang tersedia tidak seimbang[21].

b. Precision (Presisi):

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Metrik lain yang digunakan adalah precision (positive predictive value) dengan menggunakan persamaan rumus (4)[22]. Precision seharusnya idealnya bernilai 1 (tinggi) untuk sebuah pengklasifikasi yang baik, dengan mencapai nilai 1. ketika pembilang dan penyebutnya sama, yaitu TP = TP + FP. Hal ini juga berarti bahwa FP (False Positive) sama dengan nol. Ketika FP meningkat, nilai penyebut menjadi lebih besar dari pembilang, dan nilai precision akan turun[21].

c. Recall (Sensitivitas):

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Recall (true positive rate) pada persamaan rumus (5) lebih mengarah kepada persentase total hasil relevan yang diklasifikasikan dengan benar oleh algoritma yang digunakan pada model machine learning. Ketika FN meningkat, nilai penyebut menjadi lebih besar dari pembilang, dan nilai recall akan turun[21].

d. F1-Score:

$$F1\text{-}Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (6)$$

Dalam klasifikasi yang baik, dibutuhkan metrik F1-score pada persamaan rumus (6) yang menerima nilai dari precision dan recall. F1-score bernilai baik jika nilai precision dan recall mendekati nilai 1, yang berarti FP dan FN juga mendekati nilai 0, sehingga metrik F1-score merupakan metrik yang lebih baik untuk mengukur hasil dari klasifikasi pada model daripada metrik accuracy[21].

3. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan untuk mengklasifikasikan pesan digital ke dalam kategori penipuan, promosi, dan pesan normal menggunakan algoritma Naïve Bayes. Proses penelitian meliputi tahap preprocessing, proses ekstraksi fitur menggunakan metode TF-IDF dan N-Gram unigram serta penerapan seleksi fitur menggunakan metode Chi-Square.

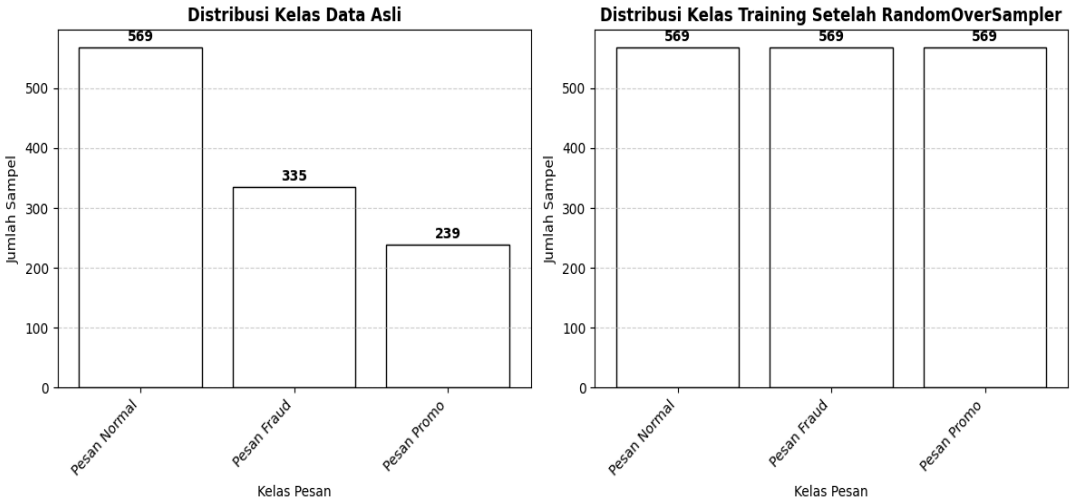
Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Hasil dari setiap tahapan menunjukkan bahwa pendekatan yang digunakan mampu meningkatkan kualitas representasi teks dan menghasilkan performa klasifikasi yang baik dalam mendeteksi pesan penipuan secara otomatis.

3.1 Random Oversampling

Setelah tahap pengumpulan data dan pelabelan awal kategori (normal, penipuan, dan promo), langkah berikutnya adalah penyeimbangan data untuk meningkatkan distribusi jumlah data pada setiap kelas. Sebelum proses penyeimbangan dilakukan, seluruh dataset terlebih dahulu dibagi menjadi data latih (*training data*) dan data uji (*testing data*) dengan rasio 80:20. Metode *Random Oversampling* kemudian diterapkan khusus pada data latih untuk meningkatkan jumlah sampel pada kelas minoritas. Proses ini dilakukan dengan menggandakan data secara acak hingga jumlah sampel pada kelas penipuan dan promo menyamai jumlah kelas mayoritas (normal) pada data latih, yaitu sebanyak 569 sampel. Penerapan metode ini bertujuan untuk mengatasi ketidakseimbangan kelas (*class imbalance*) yang dapat memengaruhi kinerja model klasifikasi, sehingga proses pelatihan (*training*) algoritma dapat berjalan secara lebih optimal.

Berdasarkan Gambar 3 Sebelum dilakukan proses Random Oversampling, distribusi label pada dataset awal menunjukkan adanya ketidakseimbangan kelas yang cukup signifikan. Dari total 1.143 data, kategori pesan normal memiliki jumlah data terbanyak yaitu 569 sampel, sedangkan kategori pesan penipuan berjumlah 335 sampel dan kategori pesan promo hanya sebanyak 239 sampel. Ketidakseimbangan distribusi data tersebut dapat menyebabkan model klasifikasi cenderung lebih dominan dalam mengenali kelas mayoritas dibandingkan kelas minoritas. Oleh karena itu, diperlukan proses penyeimbangan data agar model mampu melakukan klasifikasi secara lebih optimal pada seluruh kategori pesan. Setelah dilakukan penerapan metode Random Oversampling pada data training, distribusi jumlah data pada setiap kelas menjadi lebih seimbang. Masing-masing kategori, yaitu pesan normal, pesan penipuan, dan pesan promo, memiliki jumlah data sebanyak 569 sampel. Proses penyeimbangan data ini bertujuan untuk mengurangi bias model terhadap kelas mayoritas serta meningkatkan kemampuan model dalam mengenali pola pada kelas minoritas. Dengan distribusi data yang lebih merata, performa klasifikasi diharapkan menjadi lebih stabil dan mampu meningkatkan kualitas generalisasi model pada proses pengujian.



Gambar 3. Jumlah Dataset

3.2 Preprocessing Teks

Tahap preprocessing dilakukan untuk membersihkan dan menormalkan data pesan digital sebelum memasuki proses ekstraksi fitur TF-IDF dan klasifikasi menggunakan algoritma Naïve Bayes. Pada penelitian ini, preprocessing terdiri dari beberapa tahapan, yaitu case folding, normalisasi, filtering (*stopword removal*), dan stemming untuk menghasilkan representasi teks yang lebih terstruktur sehingga lebih mudah diproses oleh sistem..

a. Case folding

Case folding adalah tahapan *preprocessing* yang dilakukan dengan mengubah seluruh huruf pada teks menjadi huruf kecil *(lowercase)* agar format penulisan menjadi seragam dan kata yang sama tidak dikenali sebagai kata yang berbeda oleh sistem[23]. Proses ini juga sering mencakup penghilangan angka, tanda baca, dan karakter khusus lainnya agar data teks seragam dan lebih mudah dianalisis oleh mesin[24].

Tabel 5. Case Folding

Sebelum	Sesudah
GRATIS 10GB! terdekat yg blm mengganti IndosatOoredoo SIM menjadi SIM 4G, tukarkan skrg di	gratis gb terdekat yg blm mengganti indosatooredoo sim menjadi sim g tukarkan skrg di gerai kartu bagi dan nikmati gratis gb di jaringan g

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Sebelum	Sesudah
gerai kartu Bagi dan nikmati GRATIS 10GB di jaringan 4G.	toko jaya bb discount berbagai blackberry type gemini ribu torch jt dakota jta nokia samsung info klik wwwjayabbcom
TOKO JAYA BB,,, discount 50% berbagai 900 BLACKBERRY type Gemini ribu, Torch 2,5 jt, Dakota 2,8 jta NOKIA& SAMSUNG, Info klik www.jayabb.com	

Pada Tabel 5, proses *case folding* dilakukan dengan mengonversi seluruh huruf menjadi huruf kecil serta menghapus angka maupun karakter simbol tertentu, seperti, seperti "10GB" menjadi "gb" dan "50%" "900" "2,5&2,8" yang dihapus dari teks.

b. Normalization (Normalisasi)

Tahapan normalisasi diterapkan untuk mentransformasi elemen teks seperti singkatan dan istilah non-formal ke dalam bentuk baku. Hal ini bertujuan agar data tekstual yang diolah memiliki format yang standar sebelum masuk ke tahap analisis selanjutnya. Pada penelitian ini, proses normalisasi menggunakan kamus kata pada file *key_norm.csv* yang berisi pasangan kata singkatan beserta hasil normalisasinya, misalnya kata "yg" diubah menjadi "yang". Tahap ini bertujuan untuk mengurangi perbedaan variasi penulisan kata yang memiliki arti sama sehingga kualitas representasi fitur pada proses klasifikasi pesan digital dapat meningkat.

Tabel 6. Normalisasi

Sebelum	Sesudah
Yank, isikan pulsa dl di no as ini 085323048677 nanti gw gantiin.Penting bget bos krn barusan nih gw nelpon gw tgguin yach…	yank isikan pulsa dulu di nomor as ini nanti saya gantiinpenting bget bos karena barusan nih saya menelpon saya tgguin yach
dri keputusan surat tri hub honda pin pemenang bfgs mdptkan hadiah noxv jazz care info klik pin wwwtricareindonesiacom	dari keputusan surat tri hubungi honda pin pemenang bfgs mdptkan hadiah noxv jazz care informasi klik pin wwwtricareindonesiacom

Pada Tabel 6, tahap normalisasi dilakukan dengan mengonversi kata-kata tidak baku atau singkatan menjadi kata baku, seperti "dl" yang diubah menjadi "dulu", "krn" menjadi "karena", dan "hub" menjadi "hubungi", sehingga kalimat menjadi lebih sesuai dengan bentuk bahasa yang standar.

c. Filtering (*Stopword Removal*)

Filtering atau stopword removal merupakan proses penyaringan kata-kata umum yang sering muncul pada teks tetapi tidak memberikan pengaruh penting dalam proses klasifikasi. Kata-kata seperti "yang", "dan", "atau", serta kata umum lainnya dihapus karena dianggap tidak memiliki informasi yang signifikan dalam menentukan kategori pesan. Tahap ini dilakukan untuk mengurangi fitur yang tidak relevan sehingga model dapat lebih memfokuskan proses klasifikasi pada kata-kata yang memiliki pengaruh terhadap identifikasi pesan digital [25].

Tabel 7. Filtering

Sebelum	Sesudah
Aku geleh merasa sendiiri ini lipstik aku merah sekali wkkw, mana dapet dikasih ini jg	geleh sendiiri lipstik merah wkkw dapet dikasih jg
Load controller bisa di aku mad? Kmren construct coba ga bs terpaksa di dlm fungsi	load controller mad kmren construct coba ga bs terpaksa dlm fungsi

Pada Tabel 7, proses *filtering* dilakukan dengan menghilangkan kata ganti dan kata hubung yang tidak memberikan makna signifikan pada teks, contohnya "aku", "merasa", "ini", dan "di". Hal ini bertujuan agar teks menjadi lebih ringkas sehingga model dapat lebih fokus pada kata-kata yang mengandung informasi utama.

d. Stemming

Dalam tahap *stemming*, dilakukan transformasi kata berimbuhan menjadi kata dasar guna menyatukan berbagai variasi morfologi kata ke dalam satu kategori fitur. Langkah ini penting untuk meminimalkan kompleksitas fitur yang memiliki kesamaan arti dalam proses analisis data.. Pada penelitian ini, proses stemming digunakan untuk menyederhanakan kata-kata dalam pesan digital, misalnya kata "mengirim", "dikirim", dan "pengiriman" diubah menjadi kata dasar "kirim". Tahap ini bertujuan untuk mengurangi jumlah variasi kata yang tidak diperlukan sehingga proses klasifikasi dapat dilakukan dengan lebih efektif dan akurat.

Tabel 8. Stemming

Sebelum	Sesudah
Hanya HARI ini! Freedom Combo XL Cuma Rp Music (Normal Rp 149rb) 20GB 89rb streaming gak pake kuota, UNLIMITED Nelp&SMS. Daftar *123#	freedom combo xl rp music normal rp streaming gak pake kuota unlimited nelpsms daftar
SELAMAT Anda sebagai pemenang promo UNDIAN POP-MIE GETLUCKY MOBIL 1unit men-dptkan	

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Sebelum		Sesudah
NISSAN JUKE pin anda:b89c7h9 Klik:www.hadiah-pop-mie-2013.webs.com	U/INFO	selamat menang promo undi popmie getlucky mobil unit mendptkan nissan juke pin andabch uinfo klikwwwhadiahpopmiewebscom

Pada Tabel 8, proses *stemming* diterapkan untuk mengembalikan kata yang berimbuhan ke bentuk dasarnya. Sebagai contoh, kata "pemenang" diubah menjadi "menang", dan kata "UNDIAN" menjadi "undi". Hal ini dilakukan agar kata-kata memiliki bentuk dasar yang seragam, sehingga memudahkan model dalam melakukan proses analisis teks.

3.3 Extraksi Fitur TF-DF & N-Gram

Setelah melalui tahap preprocessing, data pesan digital perlu dikonversi ke bentuk numerik sehingga dapat diolah oleh algoritma *machine learning*. Pada penelitian ini digunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF), yaitu teknik pembobotan kata yang digunakan untuk merepresentasikan tingkat kepentingan suatu kata dalam dokumen. TF-IDF memberikan nilai bobot lebih tinggi pada kata yang sering muncul dalam suatu pesan namun jarang ditemukan pada pesan lainnya, sehingga kata tersebut dianggap memiliki informasi yang lebih relevan dalam proses klasifikasi. Implementasi TF-IDF pada penelitian ini dilakukan menggunakan fungsi *TfidfVectorizer* dari library *scikit-learn* dengan parameter *ngram_range = (1,1)* untuk menghasilkan fitur berbasis unigram. Pendekatan unigram digunakan untuk merepresentasikan setiap kata tunggal pada pesan digital sehingga sistem dapat mengenali kata-kata penting yang sering muncul pada kategori pesan tertentu, seperti penipuan, promosi, maupun pesan normal.

	aa	aamiiiin	aamiin	ab	abadi	abai	abbee	abdul	acara	acaratks	...	yudisium	yuk	yuks	yuni	yunit	zalora	zarkasi	zjt	zona	ztkm
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1138	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1139	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1140	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1141	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1142	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1143 rows × 3415 columns

Gambar 3. Pembobotan TF-IDF

Berdasarkan Gambar 3, hasil ekstraksi fitur menunjukkan bahwa metode TF-IDF menghasilkan matriks numerik yang merepresentasikan seluruh data pesan digital ke dalam bentuk fitur kata. Setiap fitur merepresentasikan kata-kata penting yang dipilih berdasarkan tingkat kemunculan dan relevansinya terhadap keseluruhan dataset, sehingga dapat digunakan sebagai input dalam proses klasifikasi pesan penipuan menggunakan algoritma Naïve Bayes. Dalam tahap ini, teknik TF-IDF digunakan untuk mengubah teks yang telah dipra-pemrosesan menjadi representasi numerik. Gambar 3 memperlihatkan format tabular dari 3.415 data fitur yang mewakili kata-kata unik (fitur) yang diekstrak dari corpus teks dan 1.143 baris yang mewakili dokumen atau entitas teks dalam dataset. Mayoritas nilai adalah 0.0, Hal tersebut mencerminkan bahwa sebagian besar entitas kata memiliki frekuensi kemunculan nol di dalam dokumen yang dianalisis. Ini adalah karakteristik umum data teks yang direpresentasikan dalam bentuk matriks TF-IDF. Kesimpulannya, data ini merupakan hasil transformasi teks menjadi representasi numerik berbasis TF-IDF .yang dapat langsung diterapkan dalam untuk berbagai keperluan analisis dan model pembelajaran mesin.

3.4 Seleksi Fitur Chi-Square

Setelah dilakukan ekstraksi fitur dengan pembobotan TF-IDF, langkah selanjutnya adalahmelakukan seleksi data fitur menggunakan metode Chi-Square. Chi-Square adalah metode statistik yang digunakan untuk mengukur hubungan antara kata-kata (fitur) dengan label target dalam dataset. Seleksi fitur ini dilakukan untuk menentukan kata-kata yang memiliki tingkat relevansi paling tinggi dalam membedakan setiap kelas target, seperti pesan penipuan dan bukan penipuan. Fitur yang memiliki nilai Chi-Square tinggi menunjukkan bahwa kata-kata tersebut memiliki pengaruh yang kuat dalam membedakan antara pesan penipuan dan bukan penipuan. Sebaliknya, fitur dengan nilai Chi-Square rendah dianggap kurang relevan dan dapat diabaikan. Dengan cara ini, seleksi fitur membantu mengurangi jumlah kata yang digunakan dalam model, sehingga meningkatkan efisiensi dan akurasi[7].

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

	aa	aamiiiin	aamiin	ab	abadi	abai	abbee	abdul	acaratks	account	...	yudisium	yuk	yuks	yuni	yunit	zalora	zarkasi	zjt	zona	ztkm
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1138	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1139	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1140	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1141	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1142	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

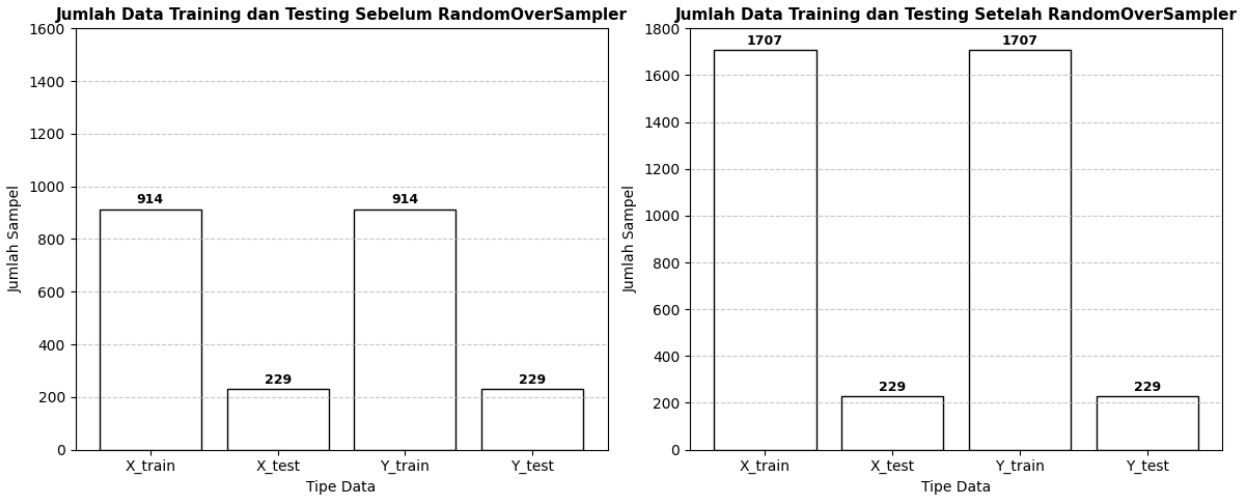
1143 rows × 3000 columns

Gambar 5. Seleksi Fitur Chi-Square

Berdasarkan Gambar 5 adalah hasil seleksi data fitur menggunakan metode Chi-Square pada dataset teks yang telah diekstraksi menggunakan TF-IDF. Dataset ini memiliki 3.000 kolom yang merupakan fitur yang telah dipilih melalui proses seleksi menggunakan Chi-Square dan 1.143 baris yang masing- masing mewakili dokumen dalam corpus. Setelah proses seleksi fitur selesai, fitur-fitur yang terpilih akan digunakan dalam tahap modeling[7].

3.5 Pelatihan Model Naive Bayes

Proses pelatihan model dilakukan setelah dataset dibagi menjadi data latih (*training data*) dan data uji (*testing data*), yang kemudian ditransformasikan ke dalam representasi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). Berdasarkan hasil pembagian awal, diperoleh sebanyak 914 data latih dan 229 data uji. Guna mengatasi ketidakseimbangan kelas pada data latih, metode *Random Oversampling* diterapkan sehingga jumlah data latih meningkat menjadi 1707 sampel, sementara data uji tetap dipertahankan sebanyak 229 sampel untuk menjaga keaslian validasi. Pada tahap pemodelan, algoritma Naïve Bayes digunakan karena memiliki efisiensi komputasi yang tinggi serta performa yang baik dalam menangani klasifikasi teks. Algoritma ini bekerja dengan mempelajari pola distribusi bobot kata dari data latih untuk menentukan probabilitas suatu pesan ke dalam kategori penipuan, promosi, atau normal.



Gambar 6. Pembagian Data Latih dan Uji

Berdasarkan Gambar 6, proses pelatihan model Naïve Bayes dilakukan menggunakan 1707 data latih yang telah diseimbangkan dan 229 data uji yang telah melalui tahap ekstraksi fitur TF-IDF. Dalam fase ini, algoritma Naïve Bayes mengidentifikasi distribusi bobot kata untuk menghitung probabilitas klasifikasi berdasarkan hubungan antar-fitur pada dataset pesan digital. Melalui proses pelatihan tersebut, model yang dihasilkan diharapkan mampu melakukan generalisasi dengan baik dan mendeteksi kategori pesan pada data uji (*testing data*) secara otomatis, cepat, dan efisien.

3.6 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa algoritma Naïve Bayes dalam mengklasifikasikan pesan digital ke dalam tiga kategori, yaitu pesan normal, pesan penipuan, dan pesan promosi. Proses evaluasi dilakukan menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*, serta divisualisasikan menggunakan *confusion matrix*.

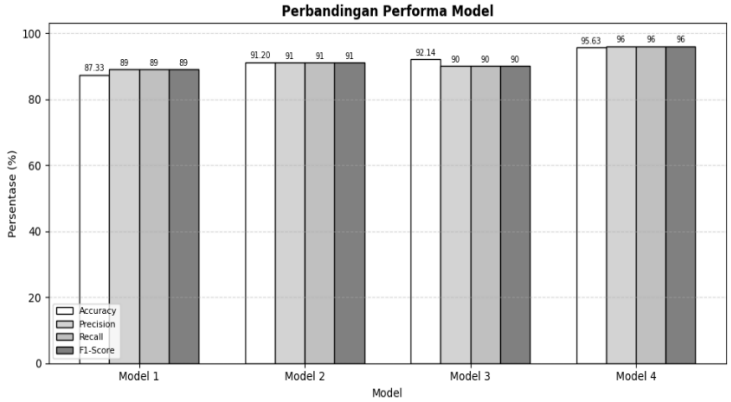
Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Berdasarkan hasil pengujian pada Tabel 9, Model 4 berhasil memperoleh nilai akurasi tertinggi sebesar 94.76%, yang menandakan bahwa sebagian besar data pengujian dapat diklasifikasikan dengan tepat oleh sistem. Selain itu, hasil *classification report* memperlihatkan bahwa performa model pada masing-masing kategori pesan berada pada tingkat yang baik dan konsisten.

Tabel. 9 Pemodelan Performa 4 Model

Model	Metode yang Digunakan	Jumlah Dataset	Accuracy	Precission	Recall	F1-Score
1	Data awal + Naive Bayes	1143	87.33%	0.89	0.89	0.89
2	Data awal + TF-IDF + Naive Bayes	1143	91.20%	0.91	0.91	0.91
3	Data awal + TF-IDF + Naive Bayes + Chi-Square	1143	92.14%	0.90	0.90	0.90
4	Random Oversampling + TF-IDF + Naive Bayes + Chi-Square	1936	95.63%	0.96	0.96	0.96

Berdasarkan Tabel 9, setiap model menunjukkan peningkatan performa seiring dengan penambahan metode yang digunakan pada proses klasifikasi. Model 1 yang hanya menggunakan algoritma Naïve Bayes memperoleh akurasi sebesar 87,33% dengan nilai precision, recall, dan F1-score masing-masing sebesar 0,89. Selanjutnya, pada Model 2 penggunaan ekstraksi fitur TF-IDF berhasil meningkatkan akurasi menjadi 91,20% dengan nilai precision, recall, dan F1-score sebesar 0,91. Pada Model 3, penambahan seleksi fitur Chi-Square mampu meningkatkan efisiensi pemilihan fitur sehingga akurasi model meningkat menjadi 92,14% dengan nilai precision, recall, dan F1-score sebesar 0,90. Sementara itu, performa terbaik diperoleh pada Model 4 yang menerapkan Random Oversampling, TF-IDF, Chi-Square, dan Naïve Bayes dengan jumlah dataset sebanyak 1707 data. Model ini berhasil mencapai akurasi sebesar 95,63% serta nilai precision, recall, dan F1-score sebesar 0,96. Hasil tersebut menunjukkan bahwa kombinasi Random Oversampling, ekstraksi fitur TF-IDF, seleksi fitur Chi-Square, dan algoritma Naïve Bayes mampu meningkatkan performa klasifikasi pesan digital secara lebih optimal dibandingkan model sebelumnya.



Gambar 7. Perbandingan Performa 4 Model

Berdasarkan Gambar 7. hasil pengujian terhadap lima model klasifikasi, diketahui bahwa setiap penambahan tahapan pengolahan data memberikan peningkatan performa model. Model 1 yang hanya menggunakan dataset asli dan algoritma Naïve Bayes memperoleh akurasi sebesar 87,33%, kemudian meningkat pada Model 2 setelah diterapkan preprocessing dan TF-IDF menjadi 91,20%. Selanjutnya, penggunaan seleksi fitur Chi-Square pada Model 3 meningkatkan akurasi menjadi 92,12%. Performa terbaik diperoleh pada Model 4 dengan jumlah dataset sebanyak 1.707 data yang mencapai akurasi sebesar 95,63%. Hasil tersebut menunjukkan bahwa penerapan Random Oversampling, ekstraksi fitur TF-IDF, dan seleksi fitur Chi-Square mampu meningkatkan performa algoritma Naïve Bayes dalam proses klasifikasi pesan digital.

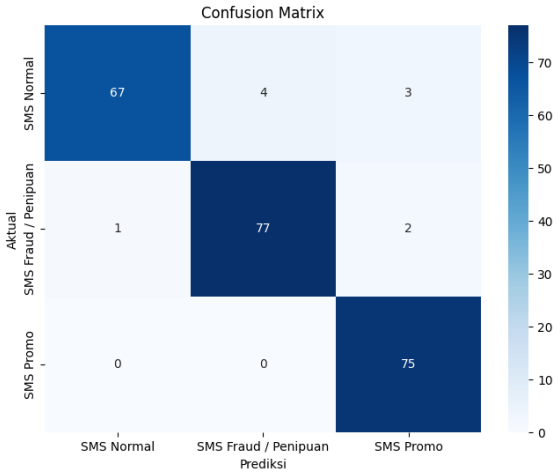
Tabel 10. Classfication Report

Label	precision	recall	F1-score	support
0 (Normal)	0.99	0.91	0.94	74
1 (Penipuan	0.95	0.96	0.96	80
2 (Promo)	0.94	1.00	0.97	75
Accuracy			0.96	229
Macro avg	0.96	0.96	0.96	229
Weigted avg	0,96	0.96	0.96	229

Berdasarkan Tabel 10, hasil pengujian model menunjukkan tingkat akurasi yang sangat baik yaitu sebesar 0,96 (atau 95,63%). Hasil tersebut membuktikan bahwa algoritma Naïve Bayes mampu melakukan klasifikasi pesan digital

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

dengan performa yang optimal pada dataset yang digunakan. Pada kategori pesan normal, model memperoleh nilai *precision* sebesar 0,99, *recall* sebesar 0,91, dan *F1-score* sebesar 0,94. Selanjutnya, kategori pesan penipuan menunjukkan performa yang stabil dengan nilai *precision* sebesar 0,95, *recall* sebesar 0,96, dan *F1-score* sebesar 0,96. Sementara itu, kategori pesan promosi mendapatkan hasil yang sangat memuaskan dengan *precision* sebesar 0,94 dan nilai *recall* sempurna sebesar 1,00, menghasilkan *F1-score* sebesar 0,97. Nilai *recall* 1,00 ini menandakan bahwa sistem berhasil mengidentifikasi seluruh pesan promosi pada data uji tanpa ada yang terlewat. Selain itu, nilai *macro average* dan *weighted average* yang secara konsisten mencapai 0,96 menegaskan bahwa kemampuan model dalam mengenali ketiga kelas pesan tersebut sangat tinggi dan seimbang. Secara keseluruhan, implementasi kombinasi TF-IDF, *N-Gram (unigram)*, seleksi fitur *Chi-Square*, dan algoritma Naïve Bayes terbukti efektif menghasilkan performa klasifikasi yang tangguh dan akurat.



Gambar 8. Confusion Matrix

Berdasarkan Gambar 8, evaluasi menggunakan *confusion matrix* menunjukkan bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan ketiga kategori pesan digital. Pada kategori Pesan Normal, model berhasil memprediksi dengan benar sebanyak 67 data dari total 74 data aktual. Pada kategori Pesan Fraud/Penipuan, model menunjukkan akurasi yang tinggi dengan mengklasifikasikan 77 data secara tepat dari total 80 data aktual, meskipun masih terdapat kesalahan minor di mana 1 data terprediksi sebagai Pesan Normal dan 2 data sebagai Pesan Promo. Sementara itu, performa paling sempurna ditunjukkan pada kategori Pesan Promo, di mana model mampu mengenali seluruh data aktual sebanyak 75 pesan secara tepat tanpa ada kesalahan klasifikasi sedikit pun ke kelas lainnya. Secara keseluruhan, hasil *confusion matrix* ini membuktikan keandalan model yang ditandai dengan dominasi nilai yang sangat tinggi pada diagonal utama matriks. Kesalahan klasifikasi yang terjadi dalam jumlah sangat kecil (yaitu pada kelas Normal dan Fraud) mengindikasikan adanya sedikit kemiripan karakteristik atau penggunaan diksi kata antar-kedua kelas tersebut. Meskipun model sudah menghasilkan performa yang superior, pengembangan lanjutan seperti optimasi bobot fitur atau eksperimen dengan arsitektur algoritma lain tetap dapat dipertimbangkan untuk mencapai hasil klasifikasi yang sepenuhnya nihil galat (*zero-error*) pada seluruh kategori pesan.

3.7 Pembahasan

Berdasarkan hasil pengujian, penelitian ini menunjukkan performa klasifikasi yang sangat baik dibandingkan beberapa penelitian terdahulu. Model Multinomial Naïve Bayes yang diterapkan berhasil memperoleh tingkat akurasi sebesar 95,63% setelah melalui tahapan preprocessing teks yang meliputi case folding, normalisasi, stopword removal, dan stemming, serta penerapan ekstraksi fitur TF-IDF dan seleksi fitur Chi-Square. Nilai tersebut menunjukkan peningkatan performa dibandingkan beberapa penelitian sebelumnya yang menggunakan metode serupa pada klasifikasi pesan spam dan penipuan. Selain itu, tingginya nilai precision, recall, dan f1-score pada setiap kategori menunjukkan bahwa model dapat mengidentifikasi karakteristik pesan normal, penipuan, dan promosi secara stabil dan konsisten. Hasil ini membuktikan bahwa kombinasi Random Oversampling, TF-IDF, Chi-Square, dan Multinomial Naïve Bayes efektif dalam meningkatkan performa klasifikasi pesan digital serta berpeluang untuk diimplementasikan pada sistem pendeteksian spam dan penipuan di lingkungan nyata.

Tabel 11. Perbandingan Penelitian

No	Peneliti	Tahun	Metode	Accuracy	F1 score
1	Sanhaji	2024	TF-IDF + Naive Bayes	91%	0.91
2	Pradana	2025	TF-IDF + Chi – Square + Naive Bayes	93%	0.93
3	Dwiprayoga	2025	BoW + TF-IDF + Naive Bayes	94.44%	0.93
4		2026	TF-IDF + SVM	92%	0.92

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

5	Pratama dan Ardianto Model yang diusulkan	2026	Random Oversampling + TF-IDF + Unigram + Chi-Square + Naïve Bayes	95.63%	0.96

Berdasarkan Tabel 11, penelitian ini memperoleh performa yang lebih baik dibandingkan beberapa penelitian terdahulu dengan akurasi sebesar 95,63% dan F1-score sebesar 0,96. Hasil tersebut menunjukkan bahwa kombinasi ekstraksi fitur TF-IDF, pendekatan N-Gram unigram, seleksi fitur Chi-Square, serta algoritma Naïve Bayes mampu meningkatkan kinerja klasifikasi pesan digital secara lebih optimal dibandingkan metode pada penelitian sebelumnya.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa model klasifikasi pesan penipuan berbahasa Indonesia menggunakan algoritma Naïve Bayes dengan kombinasi TF-IDF, N-Gram unigram, dan seleksi fitur Chi-Square mampu menghasilkan performa klasifikasi yang baik dengan tingkat akurasi sebesar 95,63%. Kontribusi utama penelitian ini adalah menghasilkan model klasifikasi pesan penipuan berbahasa Indonesia yang mengombinasikan Random Oversampling, TF-IDF, N-Gram unigram, seleksi fitur Chi-Square, dan algoritma Naïve Bayes. Kombinasi metode tersebut terbukti mampu meningkatkan performa klasifikasi dibandingkan penggunaan Naïve Bayes tanpa optimasi fitur. Selain itu, penelitian ini memberikan referensi empiris mengenai penerapan metode machine learning yang efisien dan ringan untuk mendeteksi pesan penipuan pada platform komunikasi digital berbahasa Indonesia. Penerapan Random Oversampling pada data latih berhasil membantu mengatasi ketidakseimbangan distribusi kelas sehingga model dapat mengenali kategori pesan normal, penipuan, dan promosi secara lebih seimbang. Selain itu, seleksi fitur Chi-Square terbukti mampu meningkatkan efisiensi proses klasifikasi dengan memilih fitur-fitur yang paling relevan terhadap proses prediksi. Meskipun hasil yang diperoleh cukup baik, penelitian ini masih memiliki keterbatasan dalam memahami konteks kalimat yang kompleks, ambigu, maupun hubungan antar kata yang bersifat panjang (long-range dependency). Hal tersebut disebabkan karena pendekatan TF-IDF dan N-Gram unigram hanya merepresentasikan frekuensi kemunculan kata tanpa memahami hubungan semantik serta konteks mendalam pada struktur kalimat bahasa Indonesia. Akibatnya, beberapa pesan dengan pola bahasa yang mirip antar kategori masih berpotensi mengalami kesalahan klasifikasi. Oleh karena itu, penelitian selanjutnya disarankan untuk mengembangkan metode berbasis deep learning seperti LSTM, BERT, atau Transformer yang memiliki kemampuan lebih baik dalam memahami konteks kalimat dan hubungan antar kata secara lebih mendalam. Selain itu, penelitian selanjutnya juga dapat memperluas jumlah dataset dan membandingkan berbagai teknik penanganan ketidakseimbangan data untuk memperoleh performa klasifikasi yang lebih optimal. Secara keseluruhan, sistem yang dikembangkan diharapkan dapat menjadi salah satu solusi pendukung dalam meningkatkan keamanan digital melalui deteksi pesan penipuan secara otomatis, cepat, dan akurat.

REFERENCES

[1] F. N. Azzahra, T. Rohana, R. Rahmat, and A. R. Juwita, "Penerapan Metode Naive Bayes Dalam Klasifikasi Spam SMS Menggunakan Fitur Teks Untuk Mengatasi Ancaman Pada Pengguna," *J. Inf. Syst. Res.*, vol. 5, no. 3, pp. 873–880, 2024, doi: 10.47065/josh.v5i3.5070.

[2] S. R. Prusty, B. Sainath, S. K. Jayasingh, and J. K. Mantri, "SMS Fraud Detection Using Machine Learning," *Lect. Notes Networks Syst.*, vol. 431, no. May, pp. 595–606, 2022, doi: 10.1007/978-981-19-0901-6_52.

[3] A. W. Putera, Suriati, and Y. D. Lestari, "Klasifikasi Sms Spam Menggunakan Algoritma K-Nearest Neighbor," *Jikstra*, vol. 5, no. 01, pp. 43–55, 2023.

[4] G. Sanhaji, J. Julian, and H. Syah, "WFraud Alert Sebagai Prediksi Pesan Penipuan WhatsApp Menggunakan Naïve Bayes," *J. Tekno Kompak*, vol. 18, no. 1, p. 113, 2024, doi: 10.33365/jtk.v18i1.3523.

[5] Sutriawan, Siti Mutmainnah, Teguh Ansyor Lorosae, and Sahrul Ramadhan, "Model Text Embedding dan TF-IDF+Ngram untuk Meningkatkan Kinerja Algoritma Binary Classifier pada Klasifikasi SMS Palsu," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 4, no. 1, pp. 55–64, 2025, doi: 10.53513/jursi.v4i1.10582.

[6] Lenny, "Kominfo catatkan 1.730 kasus penipuan online, kerugian ratusan triliun," Katadata.co.id. [Online]. Available: https://katadata.co.id/desysetyowati/digital/63f8a599de801/kominfo-catatkan-1730-kasus-penipuan-online-kerugian-ratusan-triliun

[7] A. P. Pradana, A. M. Syarif, I. N. Dewi, and C. Irawan, "Kombinasi naive bayes dan chi-square untuk identifikasi sms penipuan," *IRCS Integr. Res. Comput. Sci.*, vol. 1, no. 1, pp. 1–22, 2025.

[8] I. K. Dwiprayoga and M. A. Raharja, "Komparasi Ekstraksi Fitur BoW dan TF-IDF untuk Klasifikasi SMS Menggunakan Naive Bayes," *Jnatia J. Nas. Teknol. Inf. dan Apl.*, vol. 3, no. 2, pp. 247–254, 2025.

[9] T. Informatika, F. Teknik, U. Nusantara, and P. Kediri, "Klasifikasi Pesan SMS Menggunakan Metode TF-IDF dan Support Vector Machine," *Semin. Nas. Teknol. Dan Sains*, vol. 5, pp. 151–160, 2026, doi: 10.29407/9y2vn411.

[10] M. I. U. Rosyidi and N. Rochmawati, "Teknik Bagging Pada Algoritma Klasifikasi Decision Tree dan SVM Untuk Klasifikasi SMS Berbahasa Indonesia," *J. Informatics Comput. Sci.*, vol. 5, no. 02, pp. 265–271, 2023, doi: 10.26740/jinacs.v5n02.p265-271.

[11] F. R. Suprihati, "Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression," *J. Sist. Cerdas*, vol. 4, no. 3, pp. 155–160, 2021, doi: 10.37396/jsc.v4i3.166.

[12] D. W. Putri and M. A. Soeleman, "Penerapan Algoritma Naïve Bayes Terhadap Sentimen Ulasan Produk Skincare Pada E-

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 278−292
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9941

Commerce Shopee," *Technol. Sci.*, vol. 7, no. 4, pp. 2218–2228, 2026, doi: 10.47065/bits.v7i4.9209.
[13] R. Dwiyansaputra, G. S. Nugraha, F. Bimantoro, and A. Aranta, "Deteksi Sms Spam Berbahasa Indonesia Menggunakan Tf-Idf Dan Stochastic Gradient Descent Classifier," *J. Teknol. Informasi, Komput. dan Apl.*, vol. 3, no. 2, pp. 200–207, 2021.
[14] A. Wahid, M. Baharulloh, R. Kahfiansyah, T. Abrilianto, A. Saifudin, and S. Mulyati, "Identifikasi SMS Spam Menggunakan Metode Naive Bayes," *Inform. Univ. Pamulang* , vol. 6, no. 3, pp. 536–539, 2021, [Online]. Available: http://openjournal.unpam.ac.id/index.php/informatika536
[15] M. A. Akbar and F. Ariany, "Komparasi Algoritma Naive Bayes dan K-Nearest Neighbor untuk Analisis Sentimen Pengguna Dompet Digital pada Google Play Store," *Technol. Sci.*, vol. 7, no. 4, pp. 2335–2348, 2026, doi: 10.47065/bits.v7i4.9285.
[16] A. Puji Astuti, S. Alam, and I. Jaelani, "Komparasi Algoritma Support Vector Machine dengan Naive Bayes Untuk Analisis Sentimen Pada Aplikasi BRImo," *J. Bangkit Indones.*, vol. 11, no. 2, pp. 1–6, 2022, doi: 10.52771/bangkitindonesia.v11i2.196.
[17] E. A. -, "Klasifikasi Penyalahgunaan Pesan singkat Menggunakan Algoritma Naïve Bayes," *Techno Xplore J. Ilmu Komput. dan Teknol. Inf.*, vol. 8, no. 1, pp. 01–07, 2023, doi: 10.36805/technoxplore.v8i1.3500.
[18] W. Mulyaningtyas, "Deteksi Email Spam Menggunakan Multinomial Naive Bayes dengan Teknik Bag of Words," *SENTRI J. Ris. Ilm.*, vol. 5, no. 2, pp. 1523–1533, 2026, doi: 10.55681/sentri.v5i2.5650.
[19] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 6, no. 2, p. 129, 2021, doi: 10.30998/string.v6i2.10133.
[20] D. Irawan, E. B. Perkasa, Y. Yurindra, D. Wahyuningsih, and E. Helmud, "Perbandingan Klassifikasi SMS Berbasis Support Vector Machine, Naive Bayes Classifier, Random Forest dan Bagging Classifier," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 10, no. 3, pp. 432–437, 2021, doi: 10.32736/sisfokom.v10i3.1302.
[21] M. B. M. Amin *et al.*, "Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, pp. 1291–1302, 2024, doi: 10.25126/jtiik.2024118121.
[22] M. Anita, B. Susanto, and L. Larwuy, "Perbandingan Metode Random Forest dan Naïve Bayes dalam Email Spam Filtering," *KUBIK J. Publ. Ilm. Mat.*, vol. 7, no. 2, pp. 88–96, 2023, doi: 10.15575/kubik.v7i2.18933.
[23] D. I. Muhammad Hairu Dzikri, Iwan Rizal Setiawan, "Penerapan Algoritma Naive Bayes Untuk Mendeteksi Penipuan Lowongan Pekerjaan," *Sist. Inf. DAN Tek. Komput.*, vol. 8, no. 1, pp. 919–926, 2014, doi: 10.51876/simtek.v9i2.392.
[24] M. Dauber Panjaitan, P. P. Adikara, and B. D. Setiawan, "Klasifikasi Spam pada Short Message Service (SMS) menggunakan Support Vector Machine," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 1, pp. 2548–964, 2024, [Online]. Available: http://j-ptiik.ub.ac.id
[25] M. Arif Sofyan, N. Rahaningsih, and R. Danar Dana, "Deteksi Sms Spam Berbahasa Indonesia Menggunakan Algoritma Support Vector Machine," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3071–3079, 2024, doi: 10.36040/jati.v8i3.9532.